Referee Report on "Impact Evaluation of the Farmer Training and Development Activity in Honduras (November 23 2011 version)"

Michael Carter
University of California, Davis

July 2012

The strongest message that emerges from this report is that we should pity the evaluation team that had multiple efforts to implement a randomized controlled trial dashed by an implementing agency (FINTRAC) that apparently could not define the eligibility rules for its program with sufficient clarity to permit the evaluation team to define eligible populations with any precision for either its intended treatment or control sample. Had the evaluation team followed its original randomized design, it would have faced such a low compliance rate amongst farms deemed eligible for the program by the evaluation team that statistical analysis would have been rendered powerless. The evaluation team thus fell back on a secondary strategy of comparing those that were ultimately selected for the program with those that were not. While this strategy is understandable and defensible, this overall process and its attendant frustrations does suggest one question and one analytical approach for the evaluation as implemented.

First, what was FINTRAC doing with its apparently extreme reliance on subjective eligibility criteria? One imagines that FINTRAC thought it could successfully 'pick winners' with its subjective criteria. If it could pick winners, and avoid wasting project resources on farmers who were not going to succeed in higher valued farming no matter what, then this is pretty important. The Nicaragua evaluation, which we will be discussing, found some evidence that roughly 25% of program participants did not benefit from the program at all. Had it been possible to ex ante identify these 'losers' ahead of time, the program could have saved substantial resources.

So the question then becomes, was FINTRAC really picking winners (could they look into the eyes of perspective farmers and see who had the right stuff?), or were they just acting in an arbitrary way? Given that this study has access to several rounds of survey data on farmers, some of whom FINTRAC deemed winners ex ante, and others whom they deemed losers, it seems it would be useful to investigate this question. One way to proceed would be to look at some kind of standard productivity measure (e.g., technical efficiency as measured, say, by frontier econometric methods) and see if indeed the winners exhibit, well, winning characteristics. If they do, then it would behoove MCC to figure and bottle whatever FINTRAC has. If not, then future programing should offer a more level playing field to farmers who seem eligible based on their objectively measured characterisitics.

The second and deeper point concerns the appropriate econometric approach given FINTRAC's approach. Irrespective of whether FINTRAC really could or could not pick winners, it seems abundantly clear that they were NOT selecting based on observables. In this case, the studies reliance on variants of propensity score matching techniques seems misplaced. As is well known, propensity score is valid when selection is based on observables. When it is not, propensity score matching does not help us compare like with like in endogenously selected treated and untreated groups.

Indeed, it seems that any of the variation on Heckman's seminal work would seem more appropriate in this context. Heckman-style estimators use the *residual* from the first stage treatment regression (the analogue to the propensity regression) and ask whether those who were surprisingly treated (i.e., a low propensity score given their observables) are hyper-productive, and that those who were surprisingly not treated (despite a high propensity score) are of lower productivity. I was thus very surprised that the evaluation did not move in this direction methodologically.

More generally, it would have been useful to move from the (sad) discussion of FINTRAC's hard to understand behavior to a more systematic discussion of the statistical properties of the selection process and the appropriate estimation techniques. As currently written, the paper simply jumps right into propensity score methods without much reflection on what was or was not going on in terms of the selection process, and what this process means for reliable estimation techniques.

While the above comments reflect my understanding of the report and its description of the selection into treatment, I must say that I was quite astounded by the figures on pages 31-32. While it is hard to have a complete discussion of these figures (as the underlying models are never laid out, discussed or rationalized), the two figures show a tremendous lack of overlapping support in the propensity scores between the treated and untreated. Since the propensity scores are obviously based on observables, how is that the untreated nearly all have propensity scores below 20%? This bunching of propensity scores makes it seem like the ineligibles were in fact easily predictable based on observables, which is the exact opposite of what the discourse on FINTRAC's murky behavior suggests. Does the propensity score regression include some kind of indicator signaling households that were in geographic zones were FINTRAC actually did not go? This might explain the concentration of predicted scores at the low end, but as already indicated, the report needs to give us a much clearer idea about what is going on with selection given its implication for proper econometric approaches.

Looking at the propensity scores (the figure on page 31), we see a pretty strong bunching of scores for the treated between 80% and 100%, again suggesting that FINTRAC was selecting based on observables. That said, below 80%, the distribution of propensity scores for the treated is nearly uniform, suggesting a lot of selection based on unobservables. I guess if the selection process had been completely random with respect to observables (i.e., based completely on unobservables), then the distributions of propensity scores for both the treated and untreated groups would be spikes at the relevant population proportion (not 50% in this case as the sample contains more untreated than treated households).

In summary on these points, I would suggest that the report more thoroughly integrate an understanding of the selection process with its selection of econometric methods.

Beyond these primary points, I felt that the readability of the paper could be greatly improved if it were reorganized a bit and presented its information in an easier to digest way. For my tastes at least, it would have been quite helpful to first see a set of descriptive statistics on treated and untreated farms, highlight differences in both outcome variables and in baseline characteristics. Starting the empirical results in this way would then allow the paper to more clearly pose the identification challenge: which of the differences in outcome variables is a causal program impact and which is a result of pre-existing differences (i.e., due to selection). A simple look at the data in this way might help elucidate the whole question of what was going on with non(?)-random selection into treatment.

Note that such a table would then allow calculation simple difference in difference estimators (which are only introduced quite late in the paper). Again depending on what was found, we could again have an informed discussed about selection mechanisms and the likely biases of the difference in difference estimates. From there it would make sense to have a more complete discussion of the selection process and then a discussion of both propensity score and Heckman-esque results.

As a final comment, I would simply note that I found the discussion of the econometrics (largely confined to the appendix) to be a bit murky. Perhaps I am alone in this regard, but I was not familiar with the term "modified regression-adjusted propensity score based estimator for the ATE." I can guess what this might mean, but a simple presentation in the body of the text (rather than an appendix that reads like pages from a STATA manual) might be helpful.

Again, my sympathies to the evaluation team for what must have been a most difficult experience. I hope my few comments might help suggest some further approaches to get a bit more a silk purse out of this project.